Estimation of Bus Arrival Times Using APC Data

Jayakrishna Patnaik, Steven Chien, and Athanassios Bladikas
New Jersey Institute of Technology

Abstract

Bus transit operations are influenced by stochastic variations in a number of factors (e.g., traffic congestion, ridership, intersection delays, and weather conditions) that can force buses to deviate from their predetermined schedule and headway, resulting in deterioration of service and the lengthening of passenger waiting times for buses. Providing passengers with accurate bus arrival information through Advanced Traveler Information Systems can assist passengers' decision-making (e.g., postpone departure time from home) and reduce average waiting time. This article develops a set of regression models that estimate arrival times for buses traveling between two points along a route. The data applied for developing the proposed model were collected by Automatic Passenger Counters installed on buses operated by a transit agency in the northeast region of the United States. The results obtained are promising, and indicate that the developed models could be used to estimate bus arrival times under various conditions.

Introduction

Public transportation planners and operators face increasing pressures to stimulate patronage by providing efficient and user-friendly service. Within the context of Intelligent Transportation Systems (ITS), Advanced Public Transportation Systems (APTS) and Advanced Traveler Information Systems (ATIS) are designed to collect, process, and disseminate real-time information to transit users via emerg-

ing navigation and communication technologies (Federal Transit Administration 1998). One of the key elements and requirements of APTS/ATIS is the ability to estimate transit vehicle arrival and/or departure times. With quickly expanding APTS-related technologies (e.g., Global Position Systems [GPS], Automatic Vehicle Location Systems [AVLS] and Automatic Passenger Counting [APC] systems), ATIS could provide timely vehicle arrival and/or departure information to en-route, wayside, and pretrip passengers for managing their journeys (Kalaputapu and Demetsky 1995; Abdelfattah and Khan 1998; Chien and Ding 1999; Dailey, Maclean, Cathey, and Wall 2001; Lin and Padmanabhan 2002).

To estimate vehicle arrival times, dynamic models may be developed using accurate data collected by new technologies (e.g., AVLS and APC). Since bus travel times between stops depend on a number of factors (e.g., geometric conditions, route length, number of intermediate stops and intersections, turning movements, incidents, etc.), stochastic traffic conditions along the route and ridership variation at stops further increase uncertainties. Thus, the goal of this study is the application of quantitative and qualitative data to develop creditable models for estimating reliable bus arrival times.

In this study, bus arrival time estimation models are developed on the basis of data collected by APC units installed in buses. One should be surprised if a new technology works exactly as intended and generates accurate data immediately after its deployment. APC systems should be no exception. Therefore, the purpose of this article is not only to develop models for estimating bus arrival times, but also to explore problems that could be encountered while processing data collected by the APC units.

Literature Review

Bus arrivals at stops in urban networks are difficult to estimate because travel times on links, dwell times at stops, and delays at intersections fluctuate spatially and temporally. The joint impact of these fluctuations may cause schedule and headway deviations as a bus moves farther from the starting terminal, thereby lengthening the average waiting time for transit users and consequently degrading the quality of service. A sound model, which could accurately estimate vehicle arrival times, would be capable of mitigating such impact to a large extent. However, developing such a model while considering the effects of time and space, varying traffic, ridership, and weather conditions is a challenging task.

AVLS, smart pager, and ATIS devices used by transit operators can provide useful information. However, these devices fall short when it comes to estimating the travel times between any two downstream stops and the arrival times at each downstream stop from the point of real-time observation. An arrival time estimation model at every downstream stop can be developed by establishing stop-to-stop travel times as a function of several significant variables (e.g., distance, number of intermediate stops, total intermediate bus halting time, and time of day) to supplement the services offered by ATIS devices (Abdelfattah and Khan 1998).

A variety of prediction models developed in previous studies were reviewed and they can be classified into univariate and multivariate forecasting models (Chien, Ding, and Wei 2002). Univariate forecasting models are designed to predict a dependent variable by describing the intrinsic relationship with its historical data mathematically. The commonly used univariate forecasting models include probabilistic estimation and time series models (Okutani and Stephanedes 1984; Stephanedes, Kwon, and Michalopoulos 1990; Delurgio 1998).

These methods usually have a short time lag while predicting in real-time. The accuracy of time series models highly relies on the similarity between real-time and historical traffic patterns. Variation of the historical average could cause significant inaccuracy in prediction results (Smith and Demesky 1995). Unlike univariate models, multivariate models can predict and explain a dependent variable on the basis of a mathematical function of a number of independent variables. The commonly-used multivariate models are regression models and state-space Kalman filtering models (Okutani and Stephanedes 1984).

Historically, regression models (both linear and nonlinear) have been popular because they are relatively easy to use, well established, comparable with other available procedures, and well suited for parameter estimation problems. Abdelfattah and Khan (1998) developed linear and nonlinear regression models with simulation data to predict bus delays and the simultaneous influence of various factors affecting delay. They obtained relatively promising results by using a microsimulation approach.

In this study, regression models were developed using data collected by APC units installed in buses to estimate vehicle arrival times at all downstream stops. These models are developed using path-based data (e.g., travel time between two stops along the route), and the travel times are defined as a function of ridership and other external independent factors. Nonetheless, regression is not the only pos-

sible estimation approach and other methods, such as artificial neural networks, have been explored (Chien, Ding, and Wei 2002).

Objective and Scope

The primary objective of this study is to develop multivariate linear regression models for estimating bus arrival times at major stops of a route in an urban network. The study examines the methodology for developing bus arrival time estimating models; the processing, analyzing, and refining of collected data; and the behavior and impact of the independent variables. The scope of this study encompasses model development and validation; analysis of variance and covariance and colinearity matrices of dependent and independent variables; and suggestions for future research on APC implementation that can benefit users and operators.

Data Collection

Previous studies (Abdelfattah and Khan 1998; Chien, Ding, and Wei 2002) indicated that bus travel times might be affected by a number of factors such as route length, ridership (which, in turn, depends on population density and major trip generators), the number of stops and intersections, and the geometry of the route. To develop a meaningful model, data collected from the study route should have substantial variability in the aforementioned factors.

In this study, data was collected from APC units installed on buses operated on a 30-mile (48 km) urban bus route by a transit agency in the northeast United States. Various data relating to trip information can be captured and recorded as the bus heads out for a trip until it reaches the final destination. After the bus reaches the garage/terminal, a centralized computer is engaged to transfer the trip data recorded by the APC to the transit agency's data center. Service along the studied route is provided by five different patterns per each direction (e.g., inbound and outbound) over different time periods. Patterns differ in terms of where the route originates/terminates, whether or not the bus visits specific locations, and the time the bus commences the trip at the origin. Because of data availability and sufficiency, only data collected from service patterns A and B were used for developing bus travel time estimation models. There are 105 intended stops in the outbound direction for each pattern. Pattern A crosses 134 intersections (89 of which are signalized) and has 24 right and 23 left turns. Twelve important stops (known as time points) have been chosen for the analysis. These time

points serve significant trip generators and are listed on the timetables distributed by the transit agency.

The study route operates 24 hours a day. Buses operating on different patterns may travel different portions of the route. The 12 time points are at identical physical locations. The scheduled run time for the route ranges from 92 to 119 minutes for the outbound trips and 78 to 113 minutes for the inbound trips. This study was based on data recorded from January through June 2002. The data contained a total of 311 trips (including 162 outbound and 149 inbound trips) and most of the data were collected during weekday operations (including 108 outbound and 96 inbound trips). In general, each trip serves more than 60 intended stops and 100 to 300 passengers. Data collected from outbound weekday trips were used to develop the proposed models for estimating bus arrival times. Table 1 illustrates the type of data collected from the APC system.

Table 1. Variables Description of APC Data

Variable	Description
Direction	Service direction (inbound or outbound)
Open Time	Recorded bus door opening time
Close Time	Recorded bus door closing time
Leg Time	Travel time between a pair of stops
Dwell Time	
On	Number of passengers boarding at a stop
Off	Number of passengers alighting at a stop
Stop Distance	Travel distance between any two consecutive stops
Distance	Cumulative distance from the origin
Pattern ID	A code associated with each pattern of the route
Stop Sequence	A unique number attached to each stop along the route
Transit Day	Date of the service
Week Day	Day of the week
Time of Day	Starting time of the trip

Data Preparation for Model Development

As mentioned previously, arrival times may be influenced by traffic conditions, ridership, number of intermediate stops, and weather condition, which, in turn, may be different depending on time of day, day of the week, and pattern ID. If one is to estimate travel times with regression models, sufficient observations (samples) should be available for developing creditable models to produce meaningful results. For example, if the 108 outbound trips were grouped by different days, time periods, and pattern IDs, the sample size in each group would not be sufficient. Furthermore, although the actual arrival time of a bus at each time point is needed, a bus may skip a stop due to the lack of demand in some time periods. Thus, the size of data in each group is further limited.

An attempt was made to include as many data as possible in the analysis, as will be described subsequently. If a door open time was available at a time point, this was the arrival time used in the analysis for that time point. The distance between each time point and the origin is assumed as fixed with respect to each pattern ID. This data was provided by the transit agency separately. The original data were further refined by generating interstop travel times, actual number of stops a bus made and the total dwell time, and number of alighting and boarding passengers between two consecutive time points where the bus actually halted during every single trip. Based on the departure time at the first time point, trips can be grouped by time period based on their dispatching time, as indicated in Table 2, where the classification and definition of the time periods and their break points were provided by the transit agency.

Table 2. Time Periods Defined by APC Data Provider

Time Period	Symbol	Description
Early Morning	Em	Trips take place between 4:00 AM - 6:59 AM
Morning Peak	Mp	Trips take place between 7:00 AM - 9:29 AM
Late Morning	Lm	Trips take place between 9:30 AM - 11:59 AM
Mid-Day	Md	Trips take place between 12:00 Noon - 12:59 PM
Early Afternoon	Ea	Trips take place between 1:00 PM - 3:29 PM
Afternoon Peak	Ap	Trips take place between 3:30 PM - 5:29 PM
Evening	Ev	Trips take place between 5:30 PM - 7:59 PM
Late Night	Ln	Trips take place after 8:00 PM or later

Buses departing from the first time point during different time periods may experience varying traffic congestion and ridership along the route and therefore deviate from their schedule. For example, during the midday, people are likely to use buses to do shopping or errands; thus, the buses may serve more stops. Also, most schools dismiss in the early afternoon, generating student ridership and school bus traffic, causing traffic congestion. On the other hand, early morning and late night trips are likely to experience the least traffic congestion. These facts signify that time period is a significant factor associated with the estimation of bus travel times.

Whenever one uses a large database, it is desirable to screen the data carefully for erroneous entries and inconsistencies, which can be generated by equipment malfunction, human errors, software bugs, and other causes. Corrections and adjustments were made to the problematic data. When a correction was impossible, erroneous records were excluded from the analysis. Data had to be corrected/eliminated primarily because of the following reasons:

1. The *Leg Time* was reported as zero. In cases where both the door open time at a subsequent stop and close time at the previous stop were available, the difference of those times was used to compute the leg time.
2. The *Stop Distance* was reported as zero. Since distance is fixed between each time point and the origin, such data were replaced by actual time point to time point distance.
3. The *Open Time* was blank. To get this time, the *Leg Time* was added to the *Close Time* of the immediately preceding stop.
4. The *Close Time* was blank. To get this time, the *Dwell Time* was added to the *Open Time* for that stop.
5. The *Stop Sequence* was reported as zero. To identify the *Stop Sequence* (and hence the time point), the cumulative distance traveled up to that stop was computed and compared with the known distance to the time points. If a time point could be identified, the record was kept; otherwise, it was dropped.
6. The *Open Time* at a subsequent stop was earlier than the *Close Time* at a previous stop. These records were dropped.
7. The *Cumulative Distance* from the origin to a particular stop was unusually longer than the average. These records were dropped.

8. Occasionally, the *Stop Distance* would be unusually high. These records were dropped.
9. Occasionally, the bus stops (there is *Dwell Time*), but there are no on or off passengers. These records were retained (particularly since *Dwell Time* is one of the independent variables used).
10. Occasionally, there is no *Dwell Time*, but there are boarding and alighting passengers. The *Dwell Time* was calculated by taking the difference between the *Door Open Time* and *Door Close Time* at that particular stop. If door time data were not available, the record was dropped.
11. *Trip-Status* (START and END) tags would show up somewhere in the middle of the trip. The tags were moved to their appropriate places.

The data were then augmented with weather information (precipitation, visibility, and wind speed) obtained from another source.

Selection of Independent Variables

The independent variables selected to develop path-based travel time estimation models were distance, number of stops, dwell times, boarding and alighting passengers, and weather descriptors. Furthermore, there was the option of generating classes of separate models for each factor (i.e., time of day, day of week, pattern ID) that can affect travel time or include that factor as an independent variable in an overall regression.

The SAS (Version 8.02) package was used to develop a set of regression models. The decision on whether a model was reasonable was based on the signs of the coefficients, values of the R-squares, t-values of the coefficients, correlation factors among the variables, and analysis of the residuals to indicate that the developed linear models would be appropriate.

The analysis of the regression results indicated that weather variables were not among the significant factors for estimating arrival times. This can be attributed to the fact that the weather data were not sufficiently detailed or that during the study period the weather variations were not significant enough to have an impact on arrival times. A general linear model was developed for the difference of actual and scheduled journey time with independent variables (e.g., week day, time period, weather) that were categorically chosen as class factors. To identify the statistical insignificance of these variables, Tukey's test (Montgomery 2001) was conducted. The p-value generated for day of the week was 0.4712, suggesting

that trips taking place on different days of the week do not contribute any measurable difference to the travel time. These results also suggest that day of the week is not significant as an independent variable. In addition, regression models generated separately for each day of the week did not exhibit differences that could be attributed to the day. On the contrary, time of day appeared to affect travel time significantly, having very small p-values (< 0.0001).

Demand-related variables (number of stops, dwell times, boarding and alighting passengers between time points) should definitely have an impact on bus travel times. However, it is obvious that they might be highly correlated to each other. For example, regressions were tested with different combinations of data, such as (1) stops, dwell time, boarding passengers, and alighting passengers; (2) stops, dwell times, and the sum of boarding and alighting passengers (i.e. number of passengers served); and (3) stops and boarding passengers. The correlation factor between number of passengers served and total dwell time within any pair of time points was as high as 0.93. Therefore, only one of these two variables was selected. Bus dwell time was chosen, as opposed to the total number of passengers served, because the count of total passengers served could be deceptive in the sense that two distinct activities (i.e., passengers boarding and alighting the bus) could be taking place simultaneously. Even so, dwell times at previous stops directly impact vehicle arrival times in further downstream stops. The regression that included all variables produced R-square values that are smaller than the ones of the model presented here. Besides distance and time period, number of stops and duration of dwell times were the most appropriate and significant independent variables with p-values of 0.15 or less. The proposed model has some independent variables that are highly correlated (e.g., dwell time and number of stops, distance and stops) and some of their coefficients do not have a very high statistical significance.

After reviewing the data, it was found that bus travel times exceed scheduled times during certain periods. The difference is greater if a bus was dispatched during the time periods of late morning, mid-day and early afternoon than during morning peak and afternoon peak. This may be due to the prohibition of street parking in the peak hours and the presence of construction activities during nonpeak periods. Due to these differences, variables associated with the time of day the trip took place (as described in Table 2), are treated as independent variables. Additionally, the pattern IDs show a unique subset of stops along the route. An analysis of numerous regression results indicated that it was best to develop separate models for each pattern.

Given the above, the general model used to estimate bus travel (and therefore arrival) time for pattern "p" from time point "i" to all downstream time points "j" is formulated as

$$T_{i,p}=b_0+b_1d_{i,j}+b_2t_{i,j}+b_3s_{i,j}+b_4Em+b_5M_p+b_6L_m+b_7M_d+b_8E_a+b_9A_p+b_{10}E_v+b_{11}L_n$$
$$\text{for } \forall\ i \text{ and } i+1 \le j \le 12$$

where:

$T_{i,p}$	is the estimated travel time from time point "i" to all downstream time points for bus pattern "p" (e.g., A, or B) (minutes)
$d_{i,j}$	is the distance between TP_i and TP_j (miles)
$t_{i,j}$	is the average of cumulative dwell time between TP_i and TP_j (minutes)
$s_{i,j}$	is the average of cumulative number of stops between TP_i and TP_j
E_m	is a binary variable that indicates Early Morning
M_p	is a binary variable that indicates Morning Peak
L_m	is a binary variable that indicates Late Morning
M_d	is a binary variable that indicates Mid-Day
E_a	is a binary variable that indicates Early Afternoon
A_p	is a binary variable that indicates Afternoon Peak
E_v	is a binary variable that indicates Evening
L_n	is a binary variable that indicates Late Night
b_0	is the intercept of the travel time estimation model
b_k	are the parameters for variables $d_{i,j}$, $t_{i,j}$, $s_{i,j}$, E_m, M_p, L_m, M_d, E_a, A_p, E_v and L_n, respectively, where k varies from 1 to 11
i	is the index of origin time points
j	is the index of destination time points

Given a pattern ID, origin time point, and time period, the proposed model can estimate the required time to travel the path to every downstream time point and thereby the vehicle arrival time at that time point. All time periods are assigned a

value of 1 if present (if the trip started in that time period), and 0 otherwise. Regressions were run both with and without intercepts. All variable notations and their associated coefficients are the same for both types of regression models. The only difference is that models having no intercepts would have their b_0 values equal to zero.

Analysis of Results

For each of the two patterns used here, it is possible to develop one path-based model to estimate bus travel time for all downstream time points from a given starting time point. It is not possible to present the results of all models in this article. A sample of path-based models with intercepts for all possible origins of Pattern A is shown in Table 3. Conversely, Table 4 presents all path-based models of Table 3 but with no intercepts. Using the same methodology, all potential models for Pattern B were also developed but are not shown here.

The models were developed using the stepwise regression method. Variables having significance level values more than 0.15 were considered to be insignificant and, hence, were not included in the model. As shown in Tables 3 and 4, the R-square values obtained ranged from 0.96 to 0.99 for all models that have intercepts and 0.99 for those that do not have any intercepts. The estimation of arrival times is largely dependent upon the travel distance between a pair of time points. This distance was provided by the transit agency and is constant for all trips. Consequently, this results in high R-square values for all models developed. The overall p-values obtained for all models of both Patterns A and B is <0.0001. The parameter estimates of morning peak, evening, and late nighttime periods are zero. This suggests that M_p, E_v, and L_n do not enter in any of the models.

Since the methodologies used to develop all models are the same, their final results are similar. Therefore, it is redundant to discuss each one of them individually and in detail. The plot of actual versus estimated bus travel time to all downstream stops for Model I from Table 4 is presented in Figure 1 and the scatterplot of the residuals in Figure 2. Both figures substantiate visually the linear relationship of the dependent variable with all independent variables that are used in the models. In addition, normal probability plots of the residuals (not shown here) indicate that the normality assumption for the distribution of residuals is not violated.

The overall model statistics for the same model (I from Table 4) are shown in the table. The stepwise selection of variables for this model was in the order of $d_{i,j}$, $s_{i,j}$, E_a, A_p, E_m, and $t_{i,j}$. Each of these independent variables as they entered into the

Figure 1. Estimated Versus Actual Travel Time (minutes)

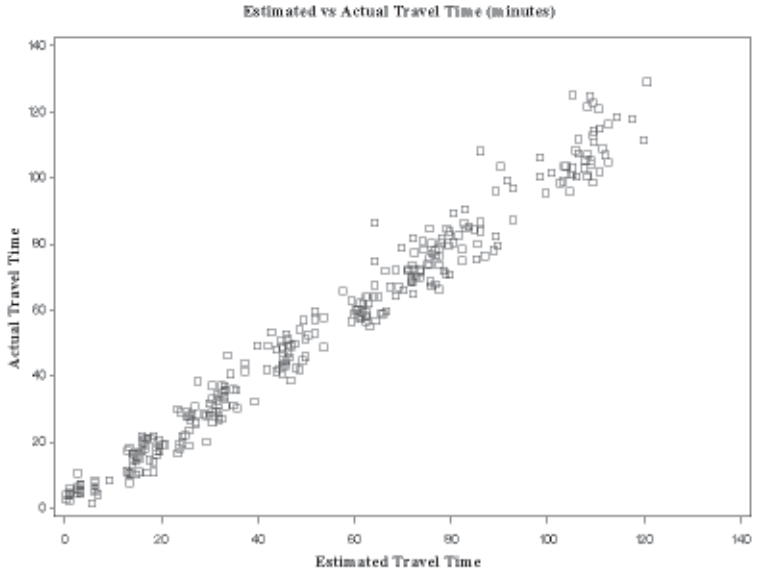
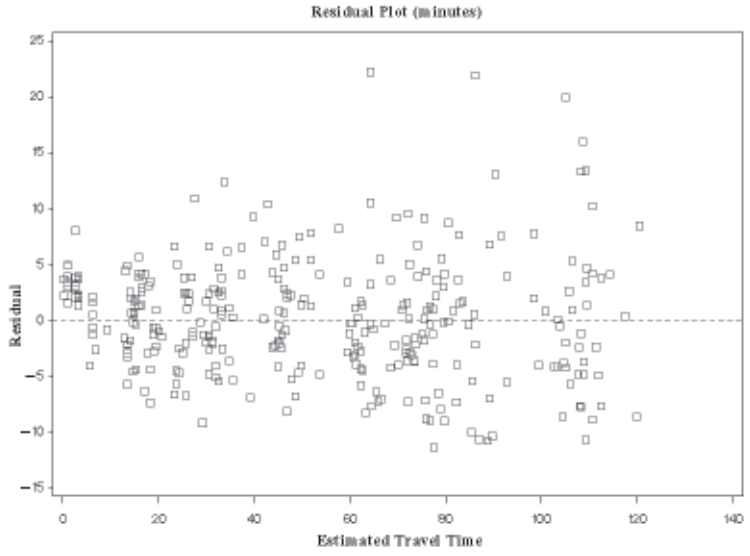


Figure 2. Residual Plot of Estimated Travel Time (minutes)



model retained their final p-values of <0.0001, 0.0914, <0.0001, 0.0084, 0.0002, and <0.0001, respectively. The summary statistics for each model are presented in Tables 3 and 4.

Table 3. Statistics of Bus Travel Time Estimation Models With Intercepts

Models	I	II	III	IV	V	VI	VII	VIII	IX
i	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9
p	A	A	A	A	A	A	A	A	A
b0	1.11	0.02	1.21	0.08	4.11	0.55	1.77	1.62	-2.86
b1	2.61	3.18	3.03	2.73	2.73	2.82	2.92	3.02	2.43
b2	0.21	0.95	0.41	0	0	0.46	0.47	0.77	5.64
b3	0.57	0	0.27	0.55	0.53	0.29	0.16	0	0
b4	-2.57	-4.39	-3.98	-3.99	-6.44	-2.78	-2.16	-2.74	0
b5	0	0	0	0	0	0	0	0	0
b6	0	-3.21	-2.37	0	-3.63	0	0	-0.77	0
b7	0	0	0	2.34	0	0	-1.13	-1.53	0
b8	2.59	0	0	5.61	0	2.61	0	0	1.98
b9	6.34	4.25	0	0	0	0	0	0	0
b10	0	0	0	0	0	0	0	0	0
b11	0	0	0	0	0	0	0	0	0
R-Sq	0.98	0.97	0.96	0.98	0.96	0.98	0.98	0.99	0.97
F value	2016.87	1372.38	1309.29	1230.32	810.44	1433.45	1123.77	858.41	307.6
RMSE	5.28	5.23	5.14	3.2	4.59	2.47	1.86	1.6	3.11
N	312	210	254	107	154	160	99	69	29
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Table 4. Statistics of Bus Travel Time Estimation Models Without Intercepts

Models	I	II	III	IV	V	VI	VII	VIII	IX
i	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9
p	A	A	A	A	A	A	A	A	A
b1	2.71	3.18	3.12	2.73	2.71	2.82	2.9	3.02	2.15
b2	0.24	0.95	0.5	0	0	0.49	0.4	0.76	5.6
b3	0.53	0	0.21	0.55	0.56	0.3	0.21	0	0
b4	-2	-4.39	-3.38	-3.94	-2.68	-2.59	-0.77	-1.11	0
b5	0	0	0	0	0	0	0	0	0
b6	0	-3.21	-1.91	0	0	0	1.31	0.86	0
b7	0	0	0	2.4	3.93	0	0	0	0
b8	3.01	0	0	5.67	3.67	2.54	1.67	1.53	0
b9	6.82	4.25	0	0	0	0	0	3.22	0
b10	0	0	0	0	0	0	0	0	0
b11	0	0	0	0	0	0	0	0	0
R-Sq	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
F value	7193.81	6142.3	6021.84	5177.83	3433.28	8314.9	4017.38	2921.03	1233.66
RMSE	5.3	5.21	5.14	3.19	4.6	2.47	1.86	1.59	3.24
N	313	211	255	108	155	161	100	70	30
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

As shown in Table 3, the travel time estimation model IX has a negative intercept of -2.86. However, this does not mean that the model will generate negative travel times. The models have positive values for the parameter estimates of variables that are reasonably significant contributors of the travel time estimation (e.g., $d_{i,j}$, $t_{i,j}$, and $s_{i,j}$), and these variables are always positive. This suggests that the estimated negative value of an intercept tends to act as an adjustor to the accuracy of a travel time estimate. Therefore, under no circumstance will a travel time estimation model generate negative travel times. Negative signs of parameter estimates for their associated indicator variables representing a specific time period can be explained similarly.

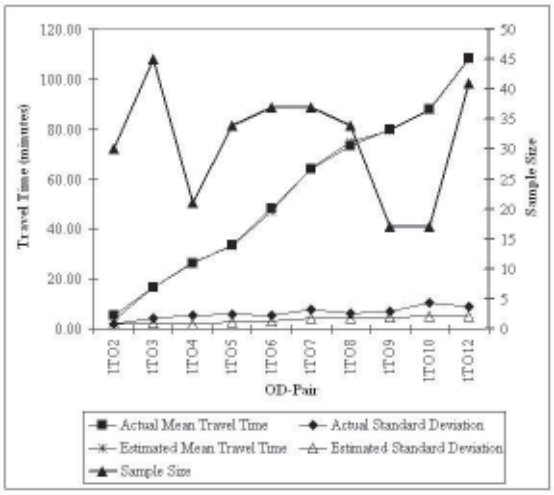
All models have a negative sign for some parameter estimate (e.g., b_4 value for variable E_m). This makes sense, because during early morning time periods, outbound buses are likely to experience less traffic congestion and, hence, shorter travel times. On the other hand, all models contained in both Tables 3 and 4 always have positive signs for parameter estimates (e.g., b_8 and b_9 for variables E_a and A_p). These results may be due to the fact that buses operating during the time periods of early afternoon and afternoon peak are expected to experience more traffic congestion and are more likely to be stopped at the signalized intersections, causing longer travel times. However, another interesting observation that can be made from these models is that some parameter estimates (e.g., b_5 for variable M_p) have either zero or negative values. This suggests that the morning peak time period either has a small or no contribution to the travel time estimation. This may be due to the fact that routes of Patterns A and B possibly experience less traffic congestion during the morning peak time period. This may be because buses are facing favorable signal timings and prohibition of street parking along the route during this time period.

A comparison of F-values of both sets of models shows that the ones that have intercepts generate smaller values than the ones that do not have any intercepts. This is consistent with the corresponding R-square values, which are a little smaller for models that have intercepts.

Data splitting or a cross validation approach (Snee 1977) is chosen for developing and then validating the models of Patterns A and B. These travel time estimation models were developed with 80 percent of the total available data for a sample size (N). The remaining 20 percent of the data were used to validate the model. Observations are chosen randomly for developing and validating the models.

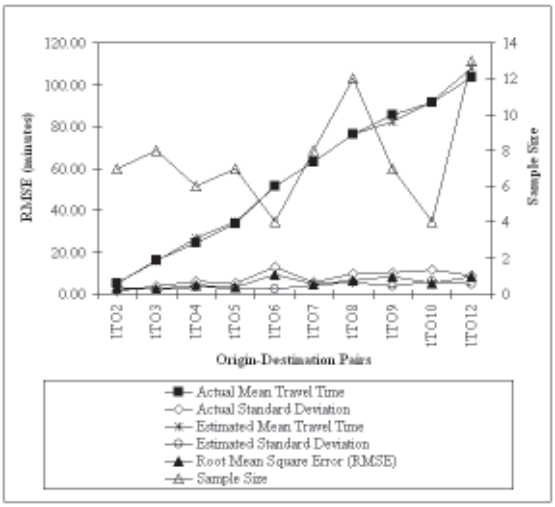
Figure 3 presents statistical descriptions of the model developed using the randomly-selected 80 percent of the total sample data available. On the other hand, Figure 4 illustrates how the 20 percent data best fits and validates the model developed by using the other 80 percent of data. The presented statistics are for the previously discussed Model I of Table 4. Means of actual versus estimated travel times for each OD pair were plotted to determine if there are any significant differences. Both Figures 3 and 4 point out that actual and estimated travel times are reasonably close to each other since the observations for model development (sample size N is equal to 313) and for model validation (sample size of 76) were randomly picked.

Figure 3. Model Development Statistics (80% of data)



As shown in Figure 4, for the OD pair TP_1-TP_6, the actual standard deviation is the highest, having a value of 12.88 minutes, while the corresponding mean actual travel time is 51.48 minutes. This may be attributed to the fact that the available sample size that was randomly chosen for this OD pair is very small and equal to 4. This explains why the root mean squared error for this OD pair is the highest (9.10) in spite of the fact that its estimated mean travel time is very close to the

Figure 4. Model Validation Statistics (20% of data)



actual mean travel time. The estimated standard deviation for this OD pair is 2.45 minutes.

The OD pair TP_1-TP_{10} has the minimum sample size of 4, as did TP_1-TP_6. But, its actual standard deviation is 11.53 minutes while its actual mean travel time is 91.49. Proportionally (as a percent of mean) this standard deviation is approximately half that of OD pair TP_1-TP_6. This can explain the smaller mean squared error value for TP_1-TP_{10} OD pair in comparison with the TP_1-TP_6 OD pair.

The OD pair TP_1-TP_{12} RMSE is 8.24 (the third highest in the sequence), in spite of its highest sample size of 13, and can be attributed to the fact that the estimated mean travel time is essentially about 5.36 minutes higher than the actual mean travel time. The estimated standard deviations of all OD pairs vary from 1.73 to 5.93 minutes, depending upon how close the downstream stops are and also what their overall sample size is. Sample size varies from 4 through 13 for all OD pairs as described.

Having mentioned all these facts, it can be concluded that the results of model validation using the 20 percent data are quite promising, suggesting that the model can be appropriately used to estimate travel times with a new set of data later. As indicated in the table and figures, the results generated by the models are

very reasonable. The plots of the estimated versus actual values indicate linear relationships. The coefficients have the anticipated signs and the adjusted R-squares are almost 0.99 for both Patterns A and B. Some models are better than others in terms of their R-squares and the statistical significance of their co-coefficients. In all cases, the mean travel time increases as we estimate travel times to farther downstream stops and so are their standard deviations. This makes sense, due to the fact that a bus is likely to encounter more and more stochastic traffic situations, causing delays as it moves farther away from the originating terminal.

On the basis of all developed models, a database can be generated that would contain parameter estimates and values of the dependent variables for the purpose of estimating the travel time at downstream stops. The transit operator would be required to input pattern ID, stop ID, and time period. Based on these inputs, the travel time estimation engine will select the appropriate model from the list of models developed to estimate the arrival times at each downstream stop. This portion of the research will commence after all models are finalized.

Conclusions and Future Research

One of the major stochastic characteristics in transit operations is that vehicle arrivals tend to deviate from the posted schedule. Poor schedule or headway adherence is undesirable for both users and operators, since it increases passenger wait/transfer times, discourages passengers from using the transit system, and degrades operating efficiency and productivity. This study developed regression models to predict bus arrival information on the basis of distance traveled, demand characteristics, and time of day. Although the available data were limited, some interpolations had to be made, and some data had to be corrected, there is no absolute certainty that some erroneous figures were not included. The initial results presented here appear to be reasonable and promising.

The methodology used for developing the travel time estimation model with APC data can be used for adjusting or planning timetables for existing or new transit routes, respectively. The developed model can be applied with ATIS to calculate and broadcast bus arrival time information at downstream stops to transit users. If a dynamic algorithm (e.g., Kalman filter) can be developed and integrated with the developed model, the accuracy of predicted bus arrival times can be greatly improved.

Another obvious comment that can be made as a result of this exercise is that one might not use indiscriminately data that are generated automatically, particularly if the system that generates them is complex and new. This is not surprising. It almost always happens, and the data quality and consistency improves rapidly with time. A good and well-known transit practitioners' example of this is the Section 15 database, which had substantial problems with the quality of its data during the first year of its release (Bladikas and Papadimitriou 1985). Therefore, the statement made here about the data quality is not meant as a criticism but as an illustration of the difficulties encountered when using new and large databases.

The data used for this study were relatively limited. The results and the models' predictive ability will certainly improve in the future when data of greater quantity and quality will be available. In the future, it may be possible to generate models for trips grouped by day, time of day, and pattern ID. Furthermore, as the ITS system deployment continues, the models could be expanded to include traffic condition variables, such as congestion and incidents, that can be automatically generated by these systems.

References

Abdelfattah, A. M., and A. M. Khan. 1998. Models for predicting bus delays. *Transportation Research Record 1623*, 8 15.

Bladikas, A., and C. Papadimitriou. 1985. A guided tour through the Section 15 maze. *Transportation Research Record 1013*, 20 27.

Chien, S., and Y. Ding. 1999. A dynamic headway control strategy for transit operations. *Conference Proceedings* (CD-ROM), 6th World Congress on ITS, Toronto, ITS Canada.

Chien, S., Y. Ding, and C. Wei. 2002. Dynamic bus arrival time prediction with artificial neural networks. *Journal of Transportation Engineering* 128 (5).

Dailey, D., S. Maclean, F. Cathey, F., and Z. Wall. 2001. Transit vehicle arrival prediction: Algorithm and large-scale implementation. *Transportation Research Record 1771*, 46 51.

DeLurgio, S. A. 1998. *Forecasting principles and applications.* New York: McGraw-Hill.

Federal Transit Administration. 1998. *Advanced Public Transportation Systems: The State of the Art, Update'98.* U.S. Department of Transportation, Washington DC.

Kalaputapu, R., and M. J. Demetsky. 1995. Application of Artificial Neural Networks and Automatic Vehicle Location Data for Bus Transit Schedule Behavior Modeling. *Transportation Research Record 1497*, 44 52.

Lin, W-H., and V. Padmanabhan. 2002. Simple procedure for creating digitized bus route information for Intelligent Transportation System applications. *Transportation Research Record 1791*, 78 84.

Montgomery, D. C. 2001. *Design and analysis of experiments.* 5th edition. John Wiley and Sons Inc.

Okutani, I. and Y. J. Stephanedes. 1984. Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research* 18B(1), 1 11.

Smith, B. L., and M. J. Demesky. 1995. Short-term traffic flow prediction: Neural network approach. *Transportation Research Record 1453*, 98 104.

Snee, Ronald. D. 1977. Validation of regression models: Methods and examples. *Technometrics* 19 (4), 15 428.

Stephanedes, Y. J., E. Kwon, and P. Michalopoulos. 1990. On-line diversion prediction for dynamic control and vehicle guidance in freeway corridors. *Transportation Research Record 1287*, 11 19.

About the Authors

Jayakrishna Patnaik (*jp42@njit.edu*) is a research assistant and masters candidate in the Department of Industrial and Manufacturing Engineering at New Jersey Institute of Technology (NJIT). He earned his bachelor's degree in mechanical engineering from Orissa University of Agriculture and Technology, India. He is a member of IIE and Alpha Pi Mu, Industrial Engineering Honors Society.

Steven I-Jy Chien (*chien@adm.njit.edu*) is an associate professor of civil engineering and has a joint appointment with the Interdisciplinary Program in Transportation at NJIT. He earned his Ph.D. degree from the University of Maryland at College Park. He is a member of ASCE, ITE, and TRB.

Athanassios Bladikas (*bladikas@adm.njit.edu*) is an associate professor of industrial and manufacturing engineering, director of the interdisciplinary program in transportation and chairperson of the Industrial and Manufacturing Engineering Department at NJIT. He earned his Ph.D. from Polytechnic University of New York and an MBA from Columbia University. He is a member of ITE, IIE, and ASEE.